On Guaranteed Optimal Robust Explanations for NLP Models

Speaker: Nicola Paoletti, King’s College London

Abstract:

We build on abduction-based explanations for machine learning and develop a method for computing local explanations for neural network models in natural language processing (NLP). Our explanations comprise a subset of the words of the input text that satisfies two key features: optimality w.r.t. a user-defined cost function, such as the length of explanation, and robustness, in that they ensure prediction invariance for any bounded perturbation in the embedding space of the left-out words. We present two solution algorithms, respectively based on implicit hitting sets and maximum universal subsets, introducing a number of algorithmic improvements to speed up convergence of hard instances. We show how our method can be configured with different perturbation sets in the embedded space and used to detect bias in predictions by enforcing include/exclude constraints on biased terms, as well as to enhance existing heuristic-based NLP explanation frameworks such as Anchors. We evaluate our framework on three widely-used sentiment analysis tasks and texts of up to 100 words from SST, Twitter and IMDB datasets, demonstrating the effectiveness of the derived explanations.

Joint work with Agnieszka Zbrzezny (University of Warmia and Mazury), Emanuele La Malfa, Rhiannon Michelmore, and Marta Kwiatkowska (University of Oxford). Appeared in IJCAI 2021.

Bio:

Nicola is a Senior Lecturer in the Department of Informatics at King's College London. In the past four years, he has been a Lecturer at the Department of Computer Science at Royal Holloway, University of London. Previously, he has been a post-doc at Stony Brook University

(USA) and University of Oxford, after an internship at Microsoft Research Cambridge (UK). He obtained his Ph.D. in Information Sciences and Complex Systems from the Universita' di Camerino (Italy).

Nicola's interests are in safety and security assurance of cyber-physical (aka autonomous) systems, or CPSs, with an emphasis on biomedical applications. His research aims to develop formal analysis methods (verification, control, and synthesis) to design CPSs that are provably correct. With CPSs increasingly incorporating machine-learning components for e.g., sensing, control and model predictions, his work also focuses on data-driven verification of CPSs, whereby formal analysis and principled learning methods come together to provide correctness guarantees and interpretability, while accounting for the uncertainty and (potential) brittleness introduced by the learning components.